Measuring Classifier Intelligence

Jim DeLeo
Department of Clinical Research Informatics
NIH Clinical Center
National Institutes of Health
Bethesda, Maryland 20892
jdeleo@nih.gov

ABSTRACT

Classifiers are seen here as systems in which input feature values are used with fitted or learned functions that produce output values which are interpreted as probabilities or fuzzy degrees of class membership, or in which output values are used with cut-off decision rules to choose bivalent class membership. Two complementary measurements for evaluating training, validation, testing, and deployment phase performances in human, mechanical, and computerized classifiers are proposed here. These measurements are derived from samples of classifier output values paired with their corresponding known probabilistic, fuzzy, or bivalent classification values. The first measurement is the area under the ROC plot. The second is the separation index newly introduced here. Both of these measurements are easy to understand and to compute. It is proposed that they be considered standard metrics for evaluating and comparing classifier intelligence.

Keywords: *classifiers, intelligence, performance metrics, intelligence metrics, area under the ROC plot, separation index, knowledge discovery from data, ensembles*

1. Introduction

The task of a human, mechanical, or computerized classifier is to use a set of values, x's, for certain particular attributes to classify an entity or event into one or more categories or classes. Classification may be bivalent, where the classifier output, y, is either negative (y=0) or positive (y=1), probabilistic where the output is a probability ($0\leq y\leq 1$) that the entity or event is associated with the bivalent positive class, or a fuzzy degree of membership ($0<y<1$) reflecting partial degree of membership in the positive class. As classification tasks become increasingly non-trivial with many attributes and highly complex nonlinear and discontinuous relationships among attribute values and classification outcome values, it may be said, in the spirit of classical artificial intelligence, that classifiers that perform well are demonstrating intelligence. Metrics are needed to measure this intelligence in order to describe and compare classifier performances. Here, two such metrics that complement each other are proposed. The first is the fairly well known area under the ROC plot. The second is a new index called the "separation index." Both metrics may be employed for bivalent, probabilistic, and fuzzy classifier outcomes. They have immediate use with present day classifiers and they have potential future use with anticipated large ensembles of autonomous intelligent classifier agents engaged in data mining for knowledge discovery purposes by means of perpetual dynamic exploration of large and expanding data bases.

2. Classifiers and Intelligent Metrics

A basic classifier is illustrated in Figure 1.

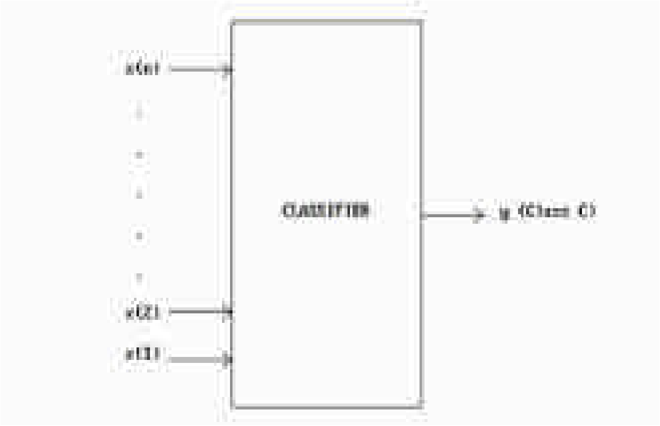

Figure 1. Illustration of a basic classifier that maps a fixed finite set of input parameter values, x's, into an output parameter value, y, where y is a measure of bivalent, probabilistic, or fuzzy classification of an entity or event with respect to some specific class, C.

Its purpose is to map a fixed finite set of input parameter values, x's, associated with an individual entity or event, E, into an output parameter value, y, where y is a measure of association of E with respect to some specific class, C associated with the output node that produces y. Basic classifiers may be designed so that y-values reflect bivalent, probabilistic,

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **AUG 2002**	2. REPORT TYPE	3. DATES COVERED **00-00-2002 to 00-00-2002**

4. TITLE AND SUBTITLE **Measuring Classifier Intelligence**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **National Institutes of Health,Department of Clinical Research Informatics,NIH Clinical Center,Bethesda,MD,20892**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Proceedings of the 2002 Performance Metrics for Intelligent Systems Workshop (PerMIS ?02), Gaithersburg, MD on August 13-15, 2002

14. ABSTRACT

see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**6**	

or fuzzy classification associations. Classifiers may have more than one output node where each node is associated with a different outcome class. The intelligence metrics described here may be applied to each separate output node. Experience suggests, however, that in polyvalent or n-class classifier applications it may be wiser to construct n basic classifiers, each having a single output node, instead of constructing one classifier with n output nodes. This is because the performance of internal mathematical "features" is reduced when these features must be "shared" in the computation of multiple outputs.

There are many mathematical methods for developing classifiers. These range from simple function fitting techniques to highly sophisticated statistical and neural network ensemble modeling. The intelligence performance metrics described here, the area under the ROC plot and the separation index, are applicable to all underlying mathematical models used in classifiers. Classifiers undergo training, learning, or fitting - terms generally used interchangeably. Classifiers also undergo validation, testing, and deployment phases. The intelligence metrics described here are intended for use in all of these phases.

3. Classifier Output Interpretation

A classifier should be designed from the outset to perform bivalent, probabilistic, or fuzzy classifications, and used in the same way throughout training, validation, testing, and deployment phases. The fundamental epistemological and mathematical differences among these three classifier types must be clearly understood at the outset of designing a classifier. These differences are based on the understanding and interpretation of the output parameter, y, and this interpretation is based on the meaning given to the class assignment data used in developing and using the classifier.

If membership data is bivalent, meaning that entities or events are perceived as belonging discretely to one or the other of the positive or negative bivalent classes, then y-values must also be interpreted as bivalent, negative or positive, usually expressed as 0 and 1 respectively. Classifiers trained with bivalent class data will often produce continuous output values for y on these intervals during all classifier phases. When this is the case, threshold decision rules are needed to force bivalent classification. If membership data is probabilistic, meaning that entities or events are perceived as belonging probabilistically to the positive pole of the bivalent classes, then output y-values may be interpreted directly as probabilities of positive class membership. For example, y= .8 could mean there is a .8 probability that the patient is a member of the bivalent set “bivalent diabetics.” If membership data is fuzzy, meaning that individual entities or events are perceived to be partly in the positive class and partly in the negative class, then output values should also be interpreted as fuzzy membership values [1]. In this case, y=.8 could mean that the patient has a .8 degree of membership in the fuzzy set "fuzzy diabetics."

Again, once a classifier system is designed to be bivalent, probabilistic, or fuzzy, it should be considered that way during training, validation, testing, and deployment phases. The interpretation of the classifier output must remain consistent throughout all of these phases.

4. The Area Under the ROC Plot

The first proposed metric of classifier intelligence is the area under the ROC plot. It is derived from ROC methodology which has origins in signal detection theory [2,3]. ROC methodology addresses forced choice bivalent classifications [4-7]. The "receiver" is a human, mechanical, or computerized agent performing the bivalent classification. "Operating characteristic" refers to the performance of the receiver. The central feature of ROC methodology is the ROC plot constructed from bivalent frequency distribution data specified as independent variable values, y, paired with dependent known classification values y_k where y_k=0 means full membership in the negative class and y_k=1 means full membership in the positive class. A basic bivalent classifier developed, for example, with neural network methodology, will have an output variable, y, with continuous values on the 0-1 interval. For purposes of applying ROC methodology, this output variable y is the independent variable which when coupled with the known classification values, y_k=0 and y_k=1 provides the data with which to compute a ROC plot. Figure 2 shows simulated bivalent frequency distributions of output values from a neural network classifier at the completion of a successful training operation. In this figure, the abscissa variable is y, the continuous neural network classifier output variable. The ordinate is the frequency at which various y values occur in both the negative class where y_k=0 (grey bars), and in the positive class where y_k=1 (black bars). It is apparent from these contrasting distributions that there are approximately the same total number of negative cases as there are positive cases. This means that the prevalence or incidence of positive events is approximately .5 in the training data. Special consideration needs to be given when there is a mismatch of prevalence in data used in training and deployment. Additional special consideration needs to be given to misclassifications cost differences, meaning differences in false positive and false negative costs. Prevalence and misclassification cost issues are very important. They have been addressed elsewhere, however more work is needed [8-12].

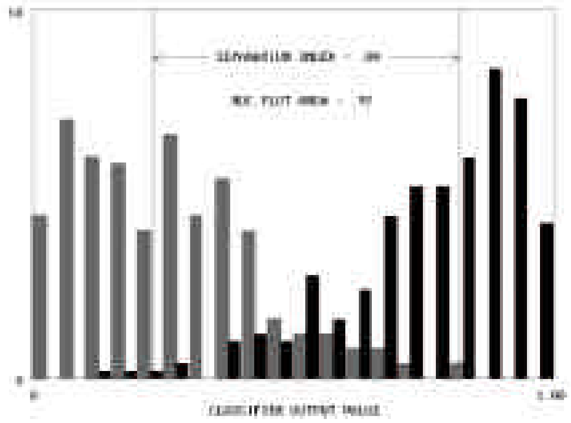

Figure 2. Bivalent frequency distributions of output values, y, from a classifier at the completion of a successful classifier neural network training operation.

The ROC plot is a plot of sensitivity versus 1-specificity as the independent variable traverses its full range. In the case of the classifier, the independent variable is y, the output of the classifier, and it ranges continuously from 0 to 1. Specificity is computed as the normalized (scaled to 1) integral of the negative distribution. Sensitivity is computed as the normalized integral of the positive distribution subtracted from 1. The ROC plot computed from the data summarized in the frequency distributions in Figure 2 is the ROC plot with an area of .97 hugging the upper left corner of the grid in Figure 3. It is displayed in this figure with 4 other ROC plots that were constructed at earlier stages in the training process. These ROC plots are empirical ROC plots meaning that they are directly computed from the y, y_k paired data. Since sensitivity and specificity are computed independently, ROC plots are, in a sense, prevalence independent.

The area under the ROC plot is readily computed by numerical integration. This area is a statistic. It is the probability that a randomly drawn event associated with the positive class will have a higher value for the independent variable, y, than a randomly drawn event associated with the negative class. In statistics it is computed as the Mann-Whitney version of the Wilcoxan statistic. The ROC plot area ranges from 0, which indicates full separation with positive cases having lower y values than the negative cases, through .5 which indicates the poorest performance (no separation of bivalent classes), to 1 which indicates the full separation of bivalent classes with positive cases having the higher y values. Classifiers with high ROC plot area values may be further differentiated in performance by means of the separation index.

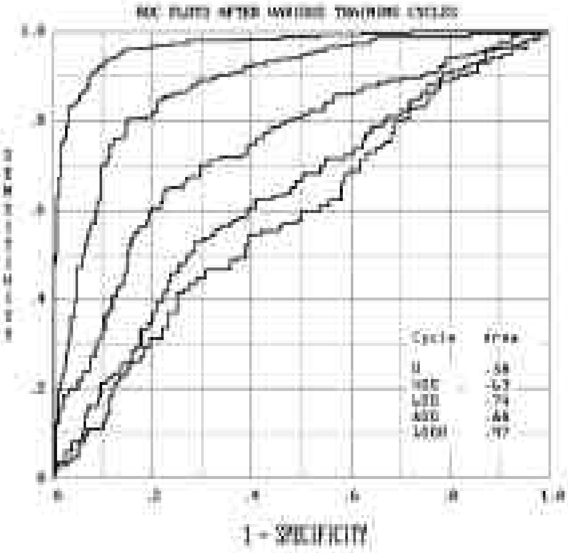

Figure 3. ROC plots from simulated neural network classifier output data after 0, 400, 600, 800, and 1000 training cycles.

5. The Separation Index

The second measure of classifier intelligence is the separation index introduced here. The separation index is a measure of the difference between the median y-values of the positive and negative frequency distributions. It is computed by first determining the median y-value for all negative cases, $n_{med,}$ and the median y-value for all positive cases, $p_{med.}$ Subtracting n_{med} from p_{med} yields a value on the -1 to +1 interval. To map this value onto the 0 to 1 interval, 1 is added and the result is divided by 2. The formula for the separation index (SI) is as follows:

$$SI = (p_{med} - n_{med} + 1) / 2 \qquad (1)$$

6. Index Complementarity

The ROC plot area and the separation index both directly measure class separation whereas other measurements used in developing classifiers generally measure the fitness of the data to the underlying function. For example, the root mean squared (RMS) error measures the square root of the sum of the squares of the differences between known, y_k, and fitted, y, outcome class values. This is clearly not a direct measure of separation. Since the task of a classifier is separation and since performing this task well requires intelligence, the ROC plot area and the separation index may be justifiably thought of as measures of intelligence since they both directly measure separation. Furthermore these indices complement one another. For example, if a ROC plot area of 1.00 indicating full separation is obtained, the separation index may be used to further differentiate training states or to compare classifiers. If poor ROC plot areas are obtained, the separation index could again indicate better or worse separation in comparing different training states or different classifiers.

7. Indices for All Classifier Phases

The ROC plot area and the separation index may be computed for the training, validation, testing, and deployment phases of classifiers for purposes of classifier evaluation and inter-classifier comparisons. Plotting values of these indices after each training cycle provides a graphical representation of the rate of intelligence development ("learning curves") during training as well as other characteristics of training, such as stalling, nonmonotonicities, and reversals. Figure 4 illustrates training plots for 1000 training cycles in the simulated experiment referred to earlier, and Table 1 contains a partial tabular listing of the data plotted in Figure 4.

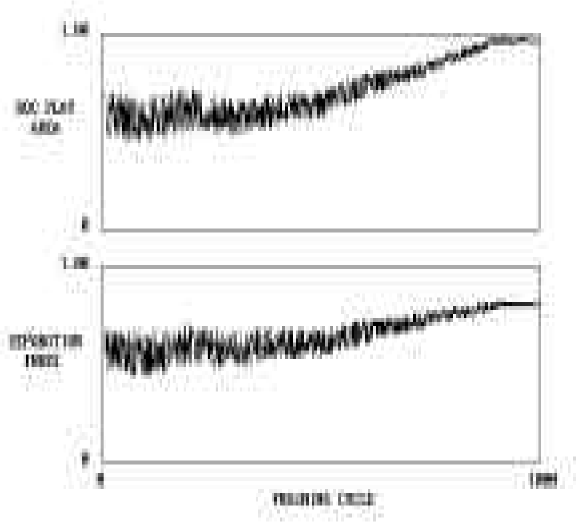

Figure 4. ROC plot area and separation index values in a simulated 1000 cycle neural network classifier training operation. (These might be thought of as "learning curves.")

What is being simulated here is the training of a neural network in which individual training cases are selected by bootstrap sampling [13-16]. This random sampling with replacement strategy is what gives rise to early higher variances tapering off to later lower variances for both indices over the training cycles. Experience suggests that this kind of sampling provides a weight jogging effect which aids in avoiding local minima entrapment. Also of note in this simulation is the observation that index values idealistically approach an asymptotic maximum. Perhaps other new metrics for intelligence could be devised for training plot features such as variance tapering and asymptotic convergence.

If validation is pursued concurrently with training, perhaps use of the area under the ROC plot and the separation index as new intelligence metrics will yield new ideas about training termination. Perhaps the intelligence metric values derived from testing data evaluated after training and validation will be considered the appropriate values to use for comparing classifiers. Monitoring intelligence metric values periodically during deployment operations would be a good way to assure that the trained classifier is holding up and that environmental data sources are not drifting too far from the original data populations associated with training, validation, and testing data sources

TRAINING CYCLE	ROC PLOT AREA	SEPARATION INDEX
0	.58	.57
100	.56	.54
200	.62	.59
300	.57	.57
400	.62	.60
500	.65	.60
600	.74	.66
700	.80	.70
800	.88	.75
900	.97	.80
1000	.97	.80

Table 1. ROC plot area and separation index values in a simulated 1000 cycle neural network classifier training operation.

8. Fuzzy and Probabilistic Membership

ROC methodology can be easily extended to include fuzzy and probabilistic classifications [17,18]. This is done by simply considering every entity or event as having relationship to both the negative and the positive bivalent poles of the class associated with the dependent variable, y_k. Let the membership association value be y_k for the positive class, and $1-y_k$ for the negative class. Thus, a fuzzy or probabilistic membership value of $y_k=.82$ corresponds with a .82 association value in the positive class and a .18 association value in the negative class. This simple generalization subsumes classical ROC methodology, because $y_k=1.00$ corresponds to an association value of 1.00 in the positive class and 0 in the negative class, and $y_k=0$ corresponds to an association value of 0 in the positive class and 1.00 in the negative class. After positive and negative class association values are determined, sensitivity and specificity values are computed from the resulting bivalent frequency distributions and the ROC plot is computed from these sensitivity and specificity values as before. The area under the ROC plot is computed by numerical integration as before or by a weighted Mann-Whitney version of the Wilcoxan statistic with tied data [18]. Likewise, the separation index is derived from the resulting bivalent frequency distributions as before.

9. Discussion

Using metrics such as those proposed here to measure classifier intelligence for evaluating and comparing classifier systems will become increasingly important in environments where large cadres of automated intelligent agents will be used in knowledge discovery from data efforts by continuously data mining large and expanding data bases. Efficient algorithms for computing intelligence metrics will be needed in these environments.

Intelligence measurement could perhaps be only one kind of measurement appropriate for evaluating intelligent systems such as classifiers. Design simplicity, computational ease, computational speed, and the capability to map knowledge produced with machine intelligence to human understandable knowledge are other important features for which metrics could be developed. Developing such measures could be an important step foreword in the progression of machine intelligence. Perhaps this step will help expand human knowledge and understanding in a more general way. By understanding intelligence and related characteristics in machines, humans may come to better understand these characteristics in humans.

10. Conclusions

Two complementary metrics, the area under the ROC plot and the separation index, have been shown to be effective measures of intelligence in all phases of classifier systems that produce bivalent, fuzzy, or probabilistic classifications. It is proposed that these metrics be standardized as measures of classifier intelligence for purposes of evaluation and comparison. The need for fast algorithms for assessing intelligence has been suggested as well as the need for measuring other attributes of classifiers and other intelligent agents, specifically attributes related to parsimony, and human knowledge derivation. The need for more work on prevalence and misclassification cost issues in classifiers has also been mentioned. It has been suggested that understanding anthropomorphized characteristics in machines may promote understanding of related characteristics in humans.

References

[1] Kosko, B., Fuzzy Thinking. New York: Hyperion, 1993.

[2] Green, D.M., Swets J.A., Signal Detection Theory and Psychophysics. New York: John Wiley & Sons, Inc., 1966.

[3] Lusted, L.B., “Signal Detectability and Medical Decision-making,” Science, vol. 171, pp. 1217-1219, 1971.

[4] DeLeo, J.M. “Receiver Operating Characteristic Laboratory (ROCLAB); Software for Developing Decision Strategies That Account for Uncertainty,” in Proceedings of the Second International Symposium on Uncertainty Modeling and Analysis, IEEE Computer Society Press, 1993, pp. 318-325..

[5] DeLeo, J.M. “The Receiver Operating Characteristic Function as a Tool for Uncertainty Management in Artificial Neural Network Decision-Making,” in Proceedings of the Second International Symposium on Uncertainty Modeling and Analysis, IEEE Computer Society Press, 1993, pp. 141-144.

[6] Lusted, L.B., “ROC Recollected,” Medical Decision Making, vol. 4, pp. 131-135, 1984..

[7] Zweig, M, Campbell, G., “Receiver Operating Characteristic (ROC) Curves: a Fundamental Tool in Clinical Medicine,” Clinical Chemistry, vol. 39, pp. 561-577, 1993.

[8] DeLeo, J., Dayhoff, J., “Medical Applications of Neural Networks: Measures of Certainty and Statistical Tradeoffs,” in Proceedings of the International Joint Conference on Neural Networks (IJCNN’01) IEEE Press, 2001, pp. 3009-3014.

[9] DeLeo, J., Rosenfeld, S., “Essential Roles for Receiver Operating Characteristic (ROC) Methodology in Classifier Neural Network Operations,” in Proceedings of the International Joint Conference on Neural Networks (IJCNN’01) IEEE Press, 2001, pp. 2730-2731.

[10] DeLeo, J.M., Rosenfeld, S.J., “Important Statistical Considerations in Classifier Systems,” in Proceedings of the Fourteenth IEEE Symposium on Computer-Based Medical Systems, IEEE Computer Society, 2001, pp. 285-293.

[11] Dayhoff, J.E., DeLeo, J.M., “Artificial Neural Networks: Opening the Black Box,” Cancer 2001 Apr 15; vol. 91 Suppl. 8, pp. 1615-1635, 2001.

[12] Remaley, A.T., Sampson, M.L., DeLeo, J.M., Remaley, N.A., Farsi, B.D., Zweig, M.H., “Prevalence-Value-Accuracy Plots: a New Method for Comparing Diagnostic tests based on Misclassification Costs,” Clinical Chemistry, vol. 45(7), 1999, pp. 934-941.

[13] Efron, B., Tibshirani, R.J., An Introduction to the Bootstrap. New York: Chapman and Hall, 1993.

[14] Efron, B., The Jackknife, the Bootstrap and Other Resampling Plans. Montpelier, Vermont: Capital City Press, 1994.

[15] Hall, P., The Bootstrap and Edgeworth Expansion. New York: Springer-Verlag, 1992.

[16] Campbell, G., DeLeo, J.M., ``Bootstrapping ROC Plots," in Computing Science and Statistics: Proceedings of the Twenty-Seventh Symposium on the Interface, 1995, pp. 420-424.

[17] DeLeo, J.M., “The Fuzzy Receiver Operating Characteristic Function and Medical Decisions with Uncertainty,” in Proceedings of the. First International Symposium on Uncertainty Modeling and Analysis, IEEE Computer Society Press, 1990, pp. 694-699.

[18] Campbell, G., DeLeo, J.M., “Fundamentals of Fuzzy Receiver Operating Characteristic (ROC) Functions,” in Computing Science and Statistics: Proceedings of the Twenty-First Symposium on the Interface, 1989, pp. 543-548.